FILE FORMATS AND STORAGE MEDIA

File Format Options

File formats quickly become obsolete as technology continually advances, making long-term preservation strategies, such as reformatting, vital. Proper advanced planning can mitigate risk and ensure that legal standards are upheld and operational requirements are met.

A file format is often described as either proprietary or nonproprietary:

- *Proprietary formats.* Proprietary file formats are controlled and supported by just one software developer. Proprietary formats, such as Microsoft Word files and WordPerfect files, carry the extension of the software in which they were created.

- *Nonproprietary formats.* These formats are supported by more than one developer and can be accessed with different software systems. For example, eXtensible Markup Language (XML) is a popular nonproprietary format.

The software in which a file is created usually has a default format, often indicated by a file name suffix, such as PDF for portable document format. Most software programs allow creators to select from a variety of formats in which to save a file, including document (DOC), Rich Text Format (RTF), and text (TXT). Some software, such as Adobe Acrobat, is designed to convert files from one format to another.

Basic file format types include the following:

1. **Text files** are most often created using word processing software. Common file formats for text files include Rich Text Format (RTF) files and Portable Document Format (PDF) files.

2. **Graphics files** store images, such as photographs and drawings, and are divided into two basic types:

   A. Vector-based files store images as geometric shapes in mathematical formulas, which allow images to be scaled without distortion.

B. Raster-based files, also referred to as bitmapped images, store images as a collection of pixels and cannot be scaled without distortion.

3. **Data files** are created in database software programs. Data files are divided into fields and tables that contain discrete elements of information. For example, a customer service database may contain customer names, addresses, and billing history fields. These fields may be organized into separate tables. Data files can be converted to a text format, but relationships among the fields and tables may be lost. Converting data files to the Comma Separated Value (CSV) file format allows aspects of the formatting and field names to be preserved.

4. **Spreadsheet files** store the value of the numbers in their cells, as well as the relationships of those numbers. For example, one cell may contain the formula that totals two other cells. Like data files, spreadsheet files are most often saved in the proprietary format of the software program in which they were created. Spreadsheet files can be exported as text files, but the value and relationship of the numbers are lost. Converting data files to the Comma Separated Value (CSV) file format allows aspects of the formatting and field names to be preserved.

5. **Video files** contain moving images, such as digitized video and animation. These files are most often created and viewed in proprietary software programs and stored in proprietary formats. At this time, best practices in the preservation of video leans towards storage on analog or digital videotape and a consistent migration plan to refresh the media.

6. **Audio files** contain sound data. These files are frequently used in customer service environments to capture audio telephone recordings as well as to track key strokes of customer service representatives, allowing these conversations to be re-created at a later time. Audio files in mp3 or other compressed formats should be converted to a preservation format. Broadcast wave format (.WAV) is considered a type of preservation format for audio files.

7. **Markup languages**, also called markup formats, contain embedded instructions for displaying or understanding the content of a file. The World Wide Web Consortium (W3C) supports these standards. Currently, eXtensible Markup Language (XML) is the most favorable format choice for long-term preservation and use of electronic records. XML, an international standard since 1998, is a human-readable, self-describing markup language

that is independent of hardware and operating systems. Because of its infrastructure-independent quality, XML is a great solution for refreshing and sharing record content. In order to use and benefit from XML, agencies must plan for certain up-front costs and time expenditures. Its structured nature, however, makes XML suitable for eventual automation and will enable the use of future open formats.

Table 1: Common File Formats

File Format Type	**Common Formats**	**Sample Files**	**Description**
Text	PDF, RTF, TXT, proprietary formats based on software (e.g., Microsoft Word)	Letters, reports, memos, e-mail messages saved as text	Created or saved as text (may include graphics)
Vector graphics	DXF, EPS, CGM	Architectural plans, complex illustrations	Store the image as geometric shapes in a mathematical formula for undistorted scaling
Raster graphics	TIFF, BMP, GIF, JPEG	Web page graphics, simple illustrations, photographs	Store the image as a collection of pixels that cannot be scaled without distortion
Data file	Proprietary to software program	Human resources files, mailing lists	Created in database software programs
Spreadsheet file	Proprietary to software program, DIF	Financial analyses, statistical calculations	Store numerical values and calculations
Video and audio files	QuickTime, MPEG	Short video to be shown on a Web site, recorded interview to be shared on CD-ROM	Contain moving images and sound
Markup languages	SGML, XML, HTML, XHTML	Text and graphics to be displayed on a Web site	Contain embedded instructions or "tags" used to transmit and display the content of a file or multiple files

File format decisions may affect electronic records management in the following ways:

- *Accessibility*. The file format must enable users to find and view the record. Records cannot be in a format that is highly compressed and easy to store if that format makes the record inaccessible.
- *Longevity*. The file format should be supported for the long-term. Proprietary software developers may not be able to ensure long-term support, thus increasing the risk of records' becoming inaccessible.
- *Accuracy*. Converted records should retain all the significant detail of the originals. The converted file should minimize data, appearance, and relationship loss.

- *Completeness*. Converted file formats should meet operational and legal objectives of existing standards for an acceptable degree of data, appearance, and relationship loss.
- *Flexibility*. The file format must meet objectives for sharing and using records. If the file format can only be read by specialized hardware and/or software, the ability to share, use, and manipulate records is limited.

Storage Options

According to the Virginia Public Records Act of the *Code of Virginia* §42.1-85, agencies are responsible for converting and migrating electronic records "as often as necessary so that information is not lost due to hardware, software, or media obsolescence or deterioration." Both agencies and localities must ensure access to electronic records for the entire length of their retention period. This means that users must be able to find, open, and read all records throughout their lifetime. Consider digital storage options that enable accessibility through migration and/or conversion of records throughout their required retention period.

In order to determine the best long-term storage medium for records, examine the current volume of stored records, along with the size of the record files and any metadata associated with them. Next, estimate projected record volume, and take into account any data access and security requirements.

There are three basic records storage options:

- *Online storage*. Records are available for immediate access and retrieval. Online storage devices include mainframe storage and network-attached storage. Online storage provides the fastest access and regular integrity checks.
- *Nearline storage*. Records are stored on media such as network-attached storage, optical disks in jukeboxes, or tapes in automated libraries. Nearline storage provides faster data access than off-line storage at a lower cost than online storage.
- *Off-line storage*. Records are stored on removable media such as magnetic tape or optical disk. Because human intervention is necessary, this option provides the slowest access.

Vital, long-term, or archival electronic records should be stored utilizing online or nearline storage options. The advantages of online and nearline storage include large storage capacities and the opportunity for data replication. Off-line storage devices are not recommended for record copies of vital, long-term, or archival records, as they are less likely to be routinely accessed and are often overlooked when systems are upgraded and electronic records are migrated to new formats. Off-line storage is recommended for backups or security copies, however, as the records can be stored off-site.

Types of Digital Storage Media

All storage media have finite life spans that are dependent on a number of factors, including manufacturing quality, age and condition before recording, handling and maintenance, frequency of access, and storage conditions. Under optimal conditions, the life expectancy of magnetic media ranges from 10 to 20 years, while optical media may last as long as 30 years. In less than ideal conditions, however, media life expectancies are significantly less.

The storage capacity of digital media is measured in bytes, the basic unit of measurement:

1,024 bytes	=	1 kilobyte (KB)
1,024 KBs	=	1 megabyte (MB)
1,024 MBs	=	1 gigabyte (GB)
1,024 GBs	=	1 terabyte (TB)
1,024 TBs	=	1 petabyte (PB)
1,024 PBs	=	1 exabyte (EB)

Access to digital information on digital media is divided into two types:

- *Sequential*. Sequentially ordered digital media requires the user to access preceding information in order to arrive at a specific point. For example, to view a specific portion of a videotape, a user must first fast-forward through the preceding portion of the videotape.
- *Random*. Some digital media allow users to access the stored information from any physical place on the media. For example, users can access any single file stored on a computer disk without having first to access all the files that precede it.

Digital media are divided into three types:

1. **Magnetic:** Electronic information is stored on computer drives, disks, or tapes by magnetizing particles imbedded in the material. Magnetic media include:

A. *Magnetic disks*, such as computer hard drives that store programs and files, are randomly accessed. Fixed disks reside permanently in a drive while removable disks are encased in plug-in cartridges, allowing for storage and transfer of data.

B. *Magnetic tape* is a sequential storage medium used for data collection, backup, and archiving. Common magnetic tape formats include Digital Audio Tape (DAT), Digital Linear Tape (DLT), and Linear-Tape Open (LTO).

2. **Optical:** Digital data is encoded by creating microscopic holes in the surface of the medium. Optical media options include:

A. *Compact Discs (CD)* can be read-only (CD-ROM), write once read many (CD-R), and rewritable (CD-RW). CDs can hold roughly 700 MB of data.

B. *Digital Versatile Discs (DVD)* are also called digital video discs. The data they store do not have to be in video form, however. DVDs can hold between 4.7 GB and 17.0 GB of data. Common types of DVDs include:

- *DVD Random Access Memory (DVD-RAM)* is a rewritable disc that provides 4.7 GB per side storage capacity.
- *DVD-R* has the same storage capacity as DVD-RAM, but can only be written to one time.
- *DVD+R* is a writable disc with 4.7 GB of storage capacity on either side.
- *DVD-RW* offer 4.7 GB per side, but can be overwritten 1,000 times. The DVD-RW technology is mainly used for video.
- *DVD+RW* is an alternative rewritable format that has a capacity of 4.7 GB per side and is used for both data and video content.

3. **Solid state:** With no moving parts, a solid state device uses electronics instead of mechanics. These devices are much faster and more reliable than magnetic and optical media. Solid state devices include:

A. A computer's BIOS (Basic Input/Output System) chip
B. PCMCIA Type I and Type II memory cards, which are used as solid-state disks in laptops
C. Flash and Dynamic Random Access Memory (DRAM-based) solid state drives

D. CompactFlash, SmartMedia, or Memory Stick, which are most often found in digital cameras

When choosing digital storage media, each option's performance characteristics must be evaluated in relation to the users' records management needs. Consider:

- How quickly users need to access the records. Some types of records require quick retrieval, while others do not.
- The volume of records that can be stored on the medium. Examine the current volume of the records and try to determine future needs.
- How long the industry will support various media options and compare those figures with the time period that records must be kept according to the approved records retention schedule. A medium that meets many needs but is not widely used or has a high risk of becoming obsolete has limited long-term value.
- How easily a given medium can be damaged or will deteriorate. A medium that deteriorates after three years might be a suitable option for records that need to be retained for only one year.
- The types of file formats a medium can store. For example, a floppy disk cannot store large graphics files, but a CD or a DVD can store graphics, text, audio files, and video files.
- The backward and forward compatibility of the digital media. A backward compatible component retains the functionality of an older component. Forward compatibility refers to the ability of the media to read information created for later versions. DVD-ROM drives are backward-compatible to CD-ROMs, but a CD-ROM drive is not forward-compatible to DVD-ROMs. This assessment will help determine how often to upgrade supporting systems, migrate and/or convert records.
- Costs and benefits of each medium, including the cost of converting and/or migrating records.